Aalto University
School of Business

ESTIMATION OF THE OPTIMAL SPEED LIMIT TO MAXIMIZE THE TRAFFIC FLOW AT GIVEN DENSITY

Master's Thesis

Zhang Wenjun
Aalto University School of Business
Information and Service Management

Fall 2021

Aalto University
School of Business

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Abstract of master's thesis

Author Zhang Wenjun

Title of thesis Estimation of the Optimal Speed Limit to Maximize the Traffic Flow at Given Density

Degree Master of Science in Economics and Business Administration

Degree programme Information and Service Management

Thesis advisor(s) Timo Kuosmanen

Year of approval 2021 **Number of pages** 62 **Language** English

Abstract

Traffic congestion on road networks has become increasingly problematic since the 1950s. During the past several decades, it has caused various serious problems. Governments and other parties have already carried out many kinds of policies to mitigate it. Variable speed limit (VSL) is one of these strategies and used on the roads to reduce traffic accidents because higher free speeds would be more likely to cause traffic congestion and to manage the traffic flows with the intention of improving capacity and then throughout the traffic.

For many years the research investigations about the relationship between traffic flow and speed limit have been being a popular topic in the transportation field. However, there still exist some conflicts on the traffic capacity and flow-density model under VSLs, some researchers proposed decreasing traffic capacity with decreasing VSL, while some induced increasing traffic capacity with decreasing VSL. The flow-density model is related to speed, traffic flow, and density, which are the main three elements of the fundamental equation in traffic flow theory. However, for all of these earlier literatures, researchers are just hypothesizing of how the flow-density diagram could look like using the simulator without considering the real-world data directly.

This thesis explored the flow-density model under VSLs based on the real-world traffic flow dataset collected from the Finnish Transport Infrastructure Agency (FTIA) with different methodology convex regression to estimate the flow-density diagram and find the optimal speed limit to maximize the traffic flow at a given density. This project refers to two remarkable results that the estimated patterns look very different from those depicted in previous theoretical speculations. Firstly, the slopes of the congestion region above the critical density reveal large seasonal differences in the flow-density model. Secondly, at lower density, a lower speed limit could keep higher traffic flow.

Keywords traffic congestion, traffic accident, variable speed limit (VSL), flow-density model, traffic flow theory, convex regression

Acknowledgements

I thank my thesis supervisor, professor Timo Kuosmanen, who allowed me to follow the topic I am interested in. Because I have never learned traffic flow theory and convex regression in system, professor Timo is so nice to help me in the related documents collection and the understanding of these professional terminologies and methodologies. In addition, professor Timo arranged enough meeting chances to discuss with me and help me find the interesting thesis topic at the earlier period and he always help me focus on these wonderful fields and remarkable results. Professor Timo always provides valuable comments on my useful findings and criticizes the meaningless parts to make sure the thesis more wonderful.

I also thanks to my family members, especially my husband Zhu Chao who accompanied me all the time during the special COVID 2019 period. It's hard to keep healthy mood and focus on the master thesis writing during this period that we only can stay at home. He also helps me collect traffic data used in the thesis, such as the speed limit in highways and the exact locations of sensors. Meanwhile, I thank my husband to help me fix my computer when it crashed this summer.

Last, but not least, I thank my friends Du Lei and Qu Ling who often encourage me when I meet some difficulties in my thesis or encounter bad mood during this period. Whenever I go to them, they always encourage that I can do it so that I can finish my thesis.

Table of Contents

List of Tables

List of Figures

1 Introduction

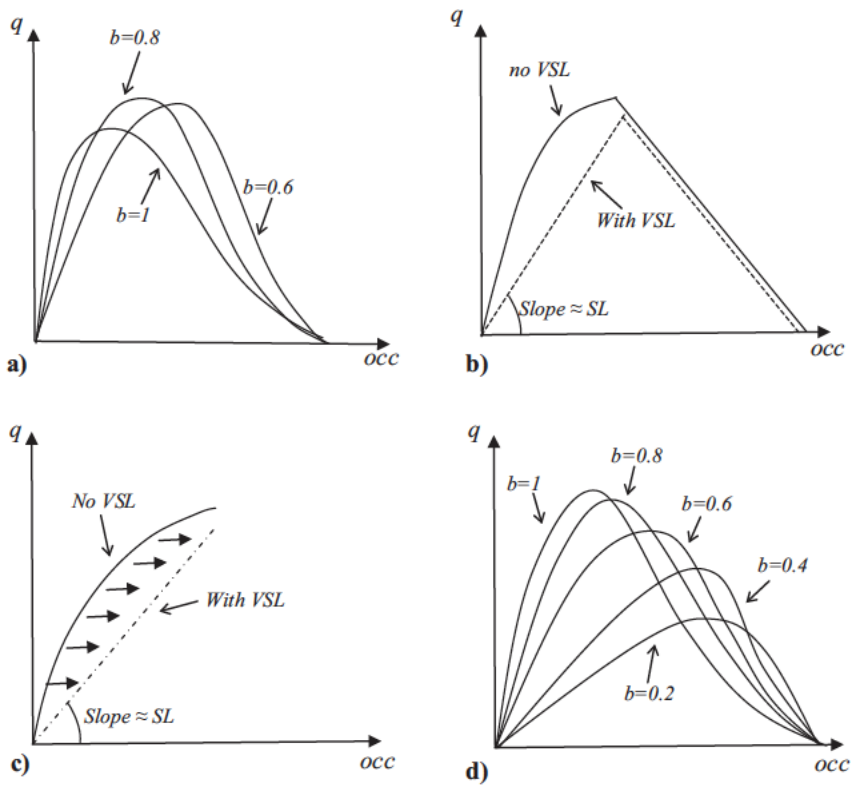
Traffic congestion on road networks has become increasingly problematic since the 1950s mentioned by Caves in 2004. During the past several decades, it has caused various serious problems: 1) the increasing opportunity cost for motorists and passengers, e.g. time-wasting, missing of meetings, time increasing of emergency vehicles; 2) the gradual deterioration of air pollution due to much more carbon dioxide emissions caused by idling, acceleration and braking; 3) much higher chance of traffic collisions due to tight spacing and constant stopping-and-going; 4) economic loss, such as the lower efficiency of delivery services, and so forth.

Based on the 2019 Urban Mobility Report published by Texas A&M Transportation Institute with cooperation from INRIX, David Schrank and Bill Eisele mention that in 2017, among 494 U.S. urban areas, congestion wastes 8.8 billion hours of extra travel time, 3.3 billion gallons of wasted fuel, and travelers making important trips have to add nearly 70 percent more travel time compared with light traffic condition to account for the effects of unexpected crashed, bad weather, special events, and other irregular congestion causes.

Traffic congestion has already influenced the daily life of human beings for many years, and governments have also already carried out many kinds of policies to mitigate it: 1) urban planning and design; 2) traffic restrictions and optimization; 3) road infrastructure investment and other effective strategies. The numeric speed limit is one of these strategies and was first proposed in 1861 in the United Kingdom for Automobiles that should be under 16km/h. The speed limit can reduce traffic accidents causing traffic congestion in 1965, the first known practice with a variable speed limit (VSL) took place on a 30 kilometer stretch of Germany motorway, with different speed limits of 60, 80, and 100 km/h indicating the changes in driving condition. For many years, research investigations about the relationship between traffic flow and speed limit have been being a popular topic in the transportation field.

The primary motivation for this thesis is to reduce the cost of traffic congestions from the VSL perspective. The objective of this thesis is to estimate the optimal point where should change the speed limit for the road. In other words, based on the density of the road, the thesis should estimate the reasonable speed limit setting to ensure a healthy traffic flow and reduce traffic accidents to relieve the congestion problem.

However, so far, most of the previous researches focus on the relationship between VSL and traffic capacity instead of the optimal speed limit setting. From Figure 1, there are four representative diagrams that present the relationship between VSL and traffic flow.



(a) Cremer (1979) based on data from Zackor (1972). Capacity increase was predicted as a result of homogenization. (b) Hegyi et al. (2005a, 2005b). The diagram under VSL is obtained as the intersection of a new free-flowing branch (according to the SL in force) and the previous diagram without control. This leads to capacity reductions for low-speed limits. (c) Papageorgiou et al. (2008). A decrease in the free flow speed is observed as a result of VSL. Observations are inconclusive regarding capacity and congested states, for which no model is proposed. (d) Carlson et al. (2010) based on data from Papageorgiou et al. (2008), propose a model with capacity reductions for lower speed limits. Note: b is the ratio between the speed limit and the free flow speed without VSL-control.

Figure 1: Existing models regarding VSL effects on the flow-occupancy diagram

Obviously, there even exist some conflicts on the traffic capacity and flow-density distributions under VSLs, some researcher proposed decreasing traffic capacity with decreasing VSL, while some induced increasing traffic capacity with decreasing VSL. And for the flow-density diagram, Rodrigo C. Carlson (2010) also mentioned that there seems to

be very limited empirical evidence and indeed no factual consensus on the potential impact of VSL on aggregate traffic behavior. In addition, Carlson pointed out that for most earlier researches, some long-held results regarding the VSL impact on the shape of the flow-density diagram are just conjectures.

What's more, there is still one hot discussion on the location of the start point of a congested state, namely the right part after the traffic flow goes to the highest points (traffic capacity) because some researchers point out that the fundamental diagram should be discontinuous. Francesc Soriguera (2017) also pointed out that there are no published results on the empirical effects of low-speed limits on freeway traffic flow, while there is common agreement that the capacity/speed limit relationship is a vital part of mainstream traffic flow control which should be further investigated. In addition, to my knowledge, there is no related work talking about the seasonal differences in the flow-density diagram.

Thus, there is an existing research gap in exploring the seasonal differences on the flow-density diagram and estimating the cross-points among VSLs, namely the optimal points instead of the traffic capacity. In order to achieve this objective, this thesis would focus on finding the seasonal differences in the flow-density diagram and the possibility of estimating the optimal speed limits for critical capacity at a given density. This project contributes to the existing base of research by answering the following research questions:

(1) Does the flow-density model keep similar critical capacity varying in the monthly change? If no, what's the reasons for these differences?
(2) Does the flow-density model keep different critical capacities varying in the speed limits? Is it possible to estimate the optimal speed limit to maximize the traffic flow at given density?

To answer these two research questions above, we collect the traffic data sets from the Väylävirasto which is the Finnish government agency that acts as the Ministry of Transport and Communications' administrative sector, process the data with the methodologies of traffic flow theory and convex regression, and analyze the results with statistical tests.

The structure of this thesis is comprised of six sections. In section 2, we will introduce the development of variable speed limits and the flow-density model (2.1) based on the finding of earlier researches. In addition, because the flow-density model is based on the traffic flow theory, in subsection 2.2, we will describe the history of traffic flow theory and the fundamental relationship among traffic flow, density, and speed. In section 3, we will

explain the methodology of convex regression we will use in this project to infer the flow rate under different speed limits in one concave case. In section 4, we will apply Finland traffic data of highway to the flow-density model. And in sections 5 and 6, the results and conclusion will be described in detail.

2 Earlier Research

In this project, we will infer the flow-density model under variable speed limits and traffic capacity curves from the traffic data including the three parameters: density, flow rate, and speed. In order to understand the work of this thesis, the readers should know these three important concepts first. For the flow-density model under variable speed limits, there are many researchers focusing on the relationships between the density and flow rate. Therefore, in the first section 2.1, it will describe the concept of variable speed limit (VSL) and the flow-density model application in historical researches. Because the flow-density model is based on the traffic flow theory, then in subsection 2.2, we will give a description about the traffic flow theory and briefly talk about how to calculate flow rate with the fundamental functions.

2.1 The VSL and Flow-density Model

Zackor, H (1972) was the first person who introduced the concept of freeway traffic control by means of VSL. He mentioned it in his book: “The first experiment was carried out in 1965 on a 30 km section of the A8 from Salzburg to Munich. The system consisted of mechanically variable message signs at a distance of 2 km, which could display speeds of 60, 80, and 100 km/h, and ‘danger zone’ and ‘accident.’ Personnel monitored traffic using video technology and manually controlled the signage. Studies reported a decrease in traffic disruptions and breakdowns, harmonization of the velocity distribution, and an increase in performance.” In the U.S. state of New Jersey in the late 1960s, the New Jersey Turnpike Authority (NJTA) also began using VSL signs in combination with variable message signs (Figure 2). Officials could adjust the speed limit according to weather, traffic conditions, and construction.

Figure 2: VSL sign in the United States in the late 1960s

Based on Zackor’s data in 1972, Cremer, M. (1979) and Zackor (1991) went on the work and proposed that VSL homogenization could improve the traffic capacity significantly, even up to 21%. Homogenization of speeds means the reduction of speed differences among vehicles and of mean speed differences among lanes. In Figure 3, the horizontal axis (o) means the occupancy of lanes, also refers to the density that is one of the

three parameters (average speed v, traffic flow q, and density k) of the fundamental diagram; the vertical axis means the traffic flow q, b = 1 indicates the ratio between the speed limit and the free-flow speed (FFS) without VSL-control. FFS describes the average that a motorist would travel if there were no congestion or other adverse conditions (such as bad weather). The highest point of each speed limit line refers to the traffic capacity, it is obvious that setting the speed limit can increase the traffic capacity significantly.

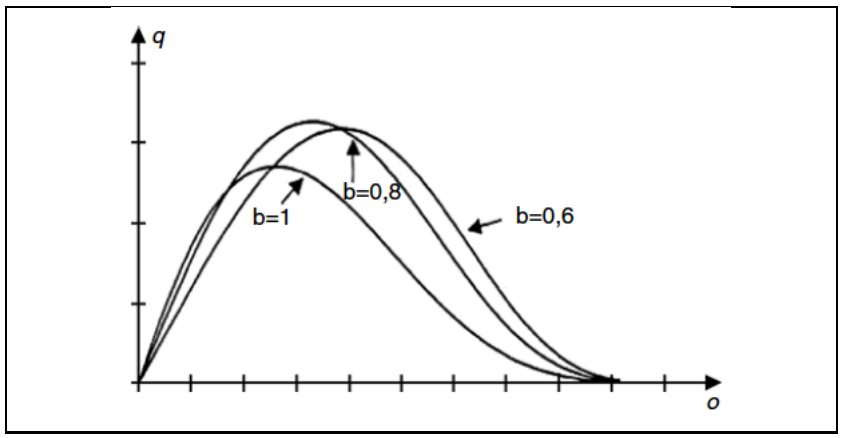
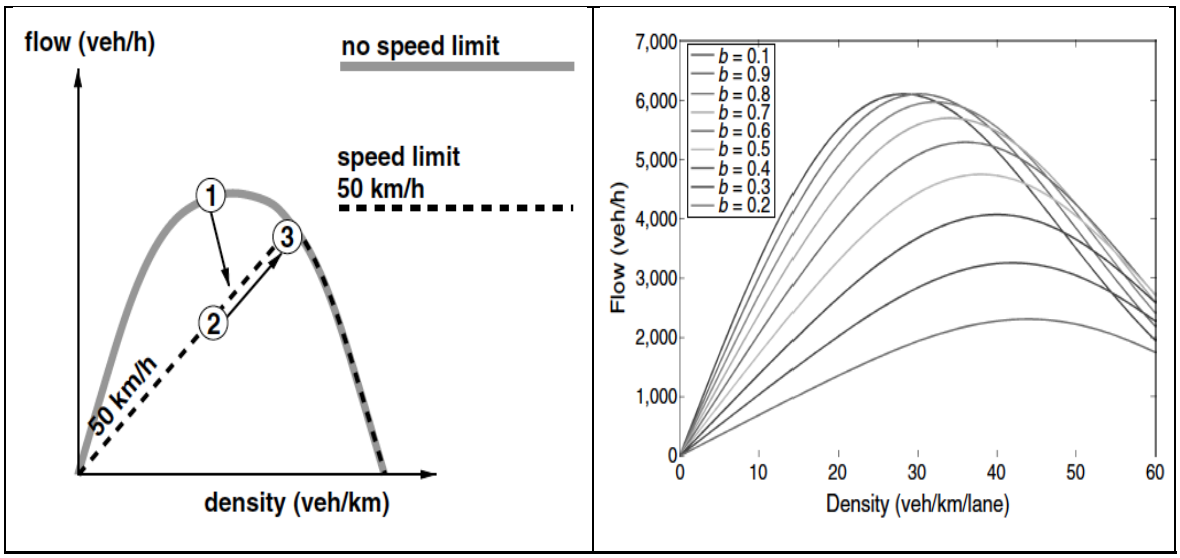


Figure 3. Cremer (1979) based on the data from Zackor (1972)

However, Andreas Hegyi (2005) and Rodrigo C. Carlson (2010) proposed opposite models that the traffic capacity would decrease when setting the speed limits. In Figure 4a, Hegyi explained that when traffic on the main road is in state 1 where is close to the traffic capacity, then it is nearly unstable and even a small flow from the on-ramp can cause a breakdown. However, speed limits could change the state from 1 to somewhere between 2 and 3 and change the shape of the fundamental diagram from the solid gray line to the dashed black line. The decreases in flow create some space for the on-ramp traffic. In Figure 4b, Carlson proposed four important implications: the free speeds are decreasing with decreasing VSL; the cross-points of VSL affected with the non-VSL curve appear near or beyond the non-VSL critical density, it is obvious that the critical densities are increasing with decreasing VSL; for b = 0.9, capacity is virtually equal as in the non-VSL case but is seen to monotonically decrease with further decreasing b values.



a) Hegyi (2005) b) Carlson (2010)

Figure 4: Distribution between density and traffic flow under VSL

Obviously, based on the previous works, there exists one controversy about traffic capacity increase as a result of VSL. Cremer and Zackor proposed that the traffic capacity would increase with decreasing speed limits, while currently Hegyi and Carlson induced that the traffic capacity would decrease with decreasing speed limits. In fact, Francesc Soriguera (2017) also pointed out that there are no published results on the empirical effects of low-speed limits on freeway traffic flow, while there is common agreement that the capacity/speed limit relationship is a vital part of mainstream traffic flow control which should be further investigated.

These conflicts and controversy among previous researches are mostly due to the difficulties in obtaining a suitable traffic database. Data is generally collected on a test corridor under specific VSL control algorithms, where different speed limits are displayed for different traffic conditions.

2.2 The Traffic Flow Theory

VSL is used on the roads to manage the traffic flows with the intention of improving capacity and then throughout the traffic. The flow-density model mentioned above is related to speed, traffic flow, and density, which are the main three elements of the fundamental equation (Greenshields, 1934) in traffic flow theory. Bruce D. Greenshields (1934) was the first person to observe traffic flows and to hypothesize a linear relationship between speed and density. He collected the data by the photographic method that the camera would take time-stamped pictures with constant intervals. These data of 1180 groups of 100 vehicles each, including not over 10 percent trucks, reveals the average free moving speed to be about 43 miles per hour on either a two or three-lane road. When the number of vehicles exceeds 400 to 600 per hour, the average speed decreases and the effect of a few slow-moving vehicles is more pronounced. Greenshields finally found the fundamental relation

$$q = kv \tag{1}$$

Where q is the traffic flow, measured in vehicles per unit of time; k is density, measured in vehicles per length of road; and v is space-mean speed, measured in length per unit of time.

This fundamental diagram defines a number of characteristics of the traffic flow. First of all, it determines the capacity or the maximum flow that can be maintained for a short period of time. The corresponding density is the so-called critical density. This capacity distinguishes two states or regimes: for densities lower than the critical density, traffic is in an uncongested state, while for higher densities, traffic is in a congested state, see Figure 5.

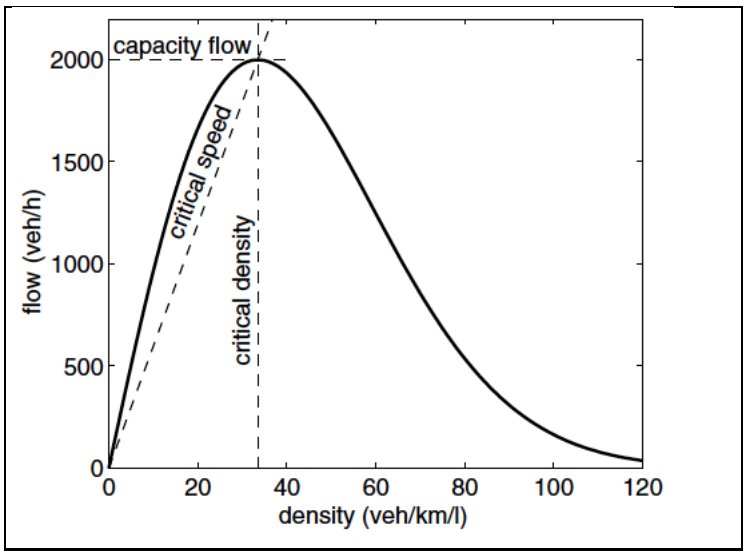


Figure 5: A typical example of the fundamental diagram

Since Greenshields, many different shapes have been proposed for the fundamental diagram. Because just from Greenshields's model, there seems to be a linear negative correlation between the number of vehicles and the average speed in a road network, however, we all know that that is too simple and it is the simplest univariate model since both the uncongested regime and the congested regime are described using the same formula.

Edie (1965) was among the first researchers to show that a discontinuous relation might be appropriate to describe traffic dynamics. He distinguished the regime of the free traffic, with the so-called free-flow capacity as maximum flow, and the congested traffic, having the so-called queue discharge rate as maximum flow.

As the thesis mentioned above, there exists one gap between free-flow capacity and speed limit capacity in Figure 4a. The same gap Edie proposed in his discontinuous fundamental diagram, the free-flow capacity is higher than the queue discharge rate, which is called the capacity drop. Said Easa (1982), Hall and Agyemang-Duah (1991) also proposed this capacity drop in their works. In this diagram figure 6, traffic has a free flow speed up to a traffic point, however, when the road situation goes to the congested state, it doesn't start the traffic capacity point, instead of one point below the traffic capacity.

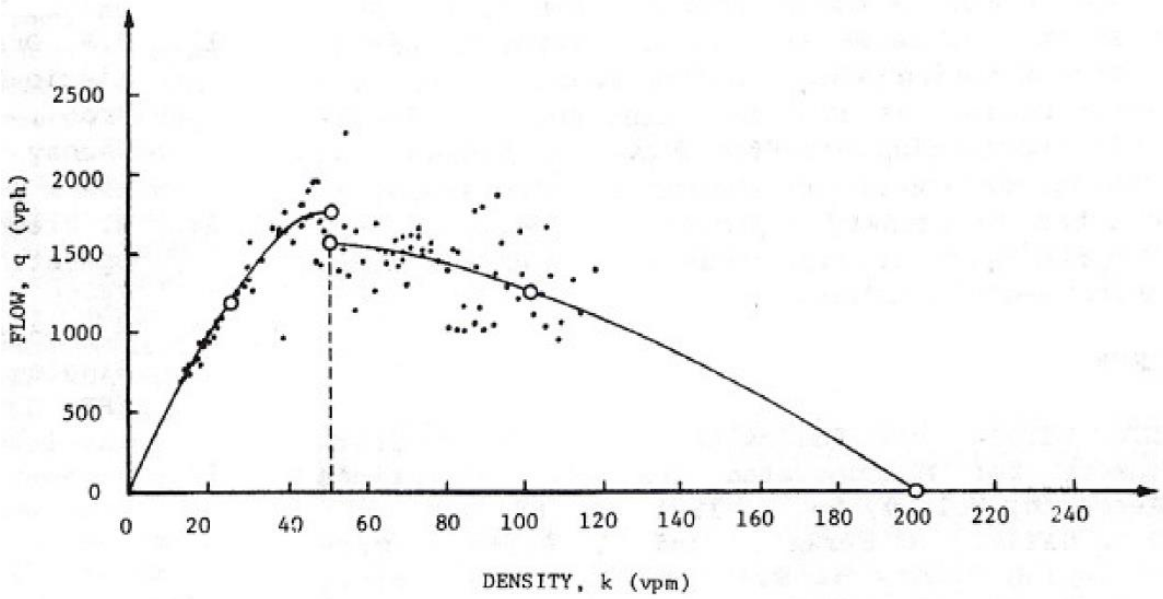


Figure 6: Discontinuous Fundamental Diagram, Said Easa (1982)

However, despite the estimated line shown in figure 6, there is no big gap or capacity drop showing the capacity discontinuity just from the data points. The inference of the flow-density diagram shown above still needs more evidence.

For more details about the fundamental relation (1), the famous traffic flow theory and its calculation of flow rate should be introduced. However, the traffic data are just observation data from the real world, we need to calculate the speed, density and flow rate firstly then draw the flow-density diagram. As for how to apply the observation data from experience to the calculate, Edie (1965) proposed one method. The total distance D traveled by all vehicles and the total time spent in the area T is calculated; the area of observation has a size of A (in units of space-time). The flow, density and speed then are calculated by

$$q = \frac{D}{A} \tag{2}$$

$$k = \frac{T}{A} \tag{3}$$

$$v = \frac{q}{k} \tag{4}$$

Suppose we have *I* vehicles passing the measurement location with a speed v_i and let us consider an area in space–time, with length λ and time interval τ. The area is A = λτ. The total distance covered is D = $\sum_i \lambda = I\lambda$; the total time spent is T = $\sum_i \lambda / vi$. Now, the speed can be computed

$$v = \frac{q}{k} = \frac{D/A}{T/A} = \frac{I\lambda}{\lambda \sum_i 1/v_i} = \frac{I}{\sum_i 1/v_i} \tag{5}$$

We can see that the fundamental diagram of traffic flow theory has already been developed into many different shapes. But as for the observation data calculation, Edie's function still could be applied in new data. Victor L. Knoop (2017) also used Edie's basic calculation method to do his simulation. Therefore, in our project, we also use this basic calculation and more details in the application will be introduced in section 4.

3 Methodology

This project applies the convex regression methodology applied in the estimation of the shape of data points in the work of Boyd & Vandenberghe（2004）, to estimate the flow-density diagram under variable speed limits in one concave case – the traffic flow theory. Obviously, the convex regression is one optimization method to get the minimization of one convex function. However, the flow-density function of the traffic flow theory shapes one concave function.

In this project, convex regression is defined as the basic method and the inference of convex regression on the concave case is regarded as the execution method when estimating the flow rate based on traffic flow theory. In section 3.1, this project will explain the mathematical basis of convex regression; and section 3.2 will apply the convex regression in the real traffic data complying with the traffic flow theory.

3.1 Convex Regression

Convex optimization is one famous method to calculate the minimization of one convex function, which is defined with continuous function and follows the rules described below. Let X be a convex set in a real vector space and let f: X → R be a function. The function is called convex function if:

$$f(tx_1 + (1-t)x_2 \leq tf(x_1) + (1-t)tf(x_2)$$
$$\forall x_1, \forall x_1 \in X, \qquad \forall t \in [0,1] \tag{6}$$

In a simple sentence to explain the convex function in mathematics, a real-valued function defined on an n-dimensional interval is called convex if the line segment between any two points on the graph of the function lies above the graph between the two points. However, these functions are defined with continuous functions instead of points of data.

The concept of using given data to automatically infer the shape of nonparametric concave or convex is from Clifford Hildreth. Hildreth mentioned in 1954, economists are frequently in the position of having fairly strong presumptions that relations among variables with which they deal satisfy certain qualitative restrictions, but they seldom have very good grounds for saying that a particular algebraic form is appropriate for representing a given relation. Before solving the economic cases, these economists often assume some properties such as the download slope of demand relation, homogeneity of a certain demand, and production relations. However, under these kinds of assumptions, the results from data actually have already been fixed in a small scope, causing the big error of simulation and prediction. Therefore, how to estimate the shape without the presumption that the given data follows some certain shape has been introduced, which is called the nonparametric regression technique. Therefore, nonparametric regression techniques that avoid strong prior assumptions about the functional form are attracting increasing attention in econometrics.

As for the concept of regression, in statistical modeling, regression is a set of statistical processes for estimating the relationships between a dependent variable (often called the outcome variable) and one or more independent variables, namely the input variables. The most common form of regression analysis is linear regression, in which a researcher finds the line (or a more complex linear combination) that most closely fits the data according to a specific mathematical criterion.

Kuosmanen in 2008 explained the canonical multiple regression model:

$$y_i = f(x_i) + \varepsilon_i, i = 1, 2, \dots, n \tag{7}$$

Where y_i is the dependent variable, f is an unknown regression function to be estimated, $x_i \in R^m$ is the vector of explanatory variables and ε_i is the idiosyncratic error term. The data set of n observations is denoted by (X,y), with y = $(y_1, \dots, y_n)' \in R^n$ and X = $(x_1, \dots, x_n) \in R^{m*n}$. So, in order to find the estimated function f close to the dataset, we need to minimize the squares of the errors. Because the error could be negative or positive, the squares of the errors will be adopted in many reteaches for estimating the differences of original data and estimated data.

After understanding how to minimize the error between given data and estimated function. The next step is to explain how to use it in convex regression. Boyd & Vandenberghe in 2004 mentioned in their textbook the simplest application to compute the least-squares fit of a convex function to given data (u_i, y_i), $i = 1, \dots, m$:

$$minimize \sum_{i=1}^{m} (y_i - f(u_i))^2 \tag{8}$$

$$subject\ to\ \ f: R^K \to R\ is\ convex, dom\ f = R^k \tag{9}$$

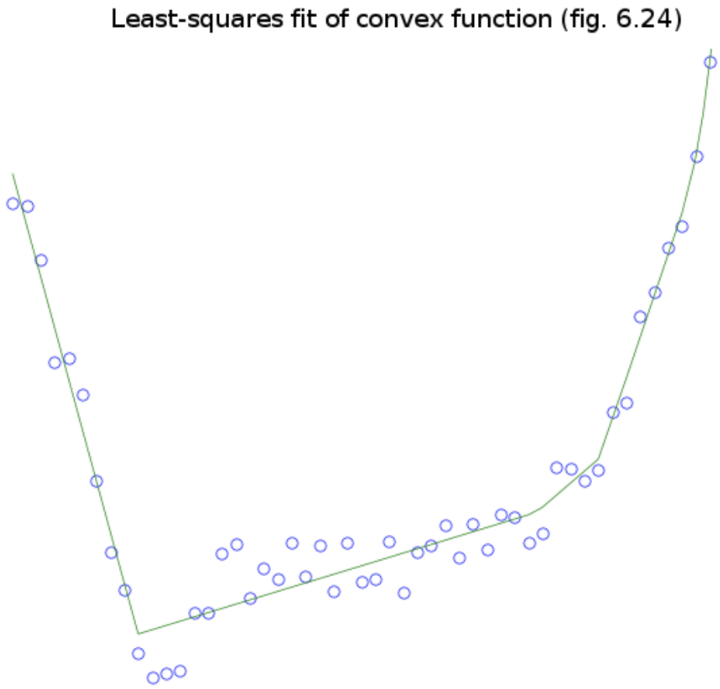
Obviously, this is an infinite-dimensional problem, since the variable is f, which is in the space of continuous real-valued functions on R^k. Given the arbitrary finite real-valued data, we can transfer this infinite-dimensional problem into the finite quadratic programming (QP) problem with variables $\hat{y} \in R^m\ and\ g_i, \dots g_m \in R^k$ as

$$minimize \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 \tag{10}$$

$$subject\ to\ \ \hat{y}_j \geq \hat{y}_i + g^T(u_j - u_i), i, j = 1, \dots, m \tag{11}$$

The inequality constraint (11) guarantees that the point set formed by the estimated $\hat{y}$ is convex.

Boyd & Vandenberghe（2004）in their textbook fitted the given data with convex set and inferred the convex function with simply interpolating linear functions between given data by maximizing variables. We can find the largest possible value for a convex function in figure 7 below.



The (piecewise-linear) function shown minimizes the sum of squared fitting error, over all convex functions. In this model, Boyd & Vandenberghe use the free CVXOPT software package for convex optimization based on the Python programming language

Figure 7: Least-squares fit of convex function to given data

3.2 Convex Regression in Traffic Flow Theory

Boyd & Vandenberghe（2004）have already applied the QP problem to fit the given data with the convex set. But the traffic flow theory follows the concave function based on earlier works. In this project, in order to estimate the capacity curve of the flow-density model with convex regression, we need to first transfer the QP problem to fit the traffic data complying with the traffic flow theory.

In the convex regression, the objective of solving the problem is to minimize the summary of squares of errors in order to get the results much closer to the model. In the flow-density model, the same objective for solving the estimation to minimize the summary of squares of errors between given data points and estimated data points. Therefore, the objective function keeps the same as the QP problem shown in function (10).

However, the constraint (11) above guarantees the convexity of the given data. In this project, we need to consider the concave property of given data. Let X be a convex set in a real vector space and let f: X → R be a function. The function is called concave function if:

$$f(tx_1 + (1-t)x_2 \geq tf(x_1) + (1-t)tf(x_2)$$
$$\forall x_1, \forall x_1 \in X, \qquad \forall t \in [0,1] \tag{12}$$

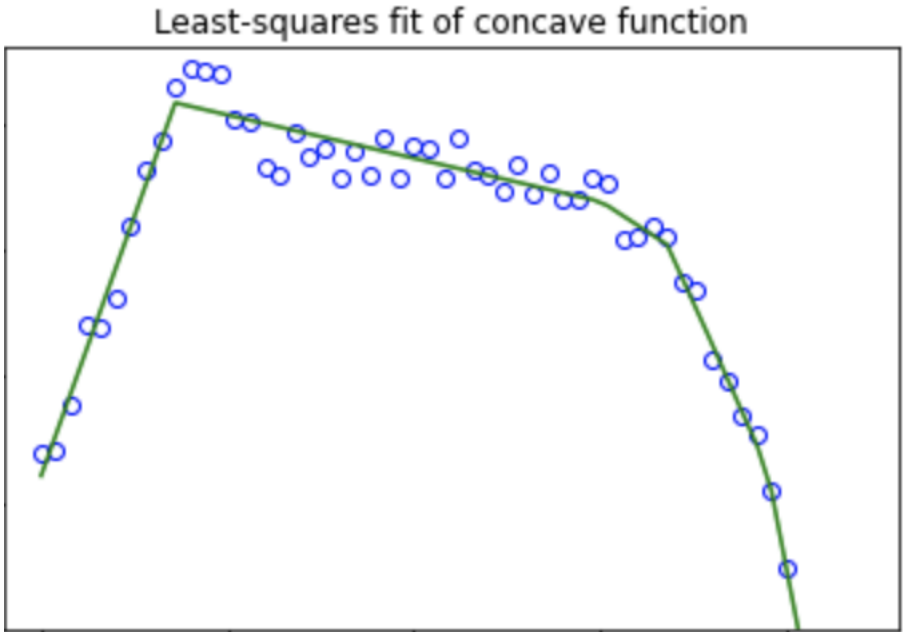
In a simple sentence to explain the concave function in mathematics, a real-valued function defined on an n-dimensional interval is called concave if the line segment between any two points on the graph of the function lies below the graph between the two pints. Therefore, the constraint (11) in convex regression should be changed to the opposite way like below.

$$subject\ to \quad \hat{y}_j \leq \hat{y}_i + g^T(u_j - u_i), i,j = 1,\dots,m \tag{13}$$

Lauren A. Hannah and David B. Dunson (2013) also mentioned that convex regression is easily extended to concave regression since a concave function is the negative of a convex function. However, they just applied convex regression in the convex dataset and proposed a new model for solving the convex regression problem, they didn't explain this theory in concave functions in detail.

In order to prove that this constraint change also fit the concave function, we trained the dataset provided by Boyd & Vandenberghe in 2004 to make sure we can get the capacity

inference curve from the changed given dataset. We change the dataset to the opposite dataset by inversing the dependent variable and change the constraints (11) with (13). Then we obtain the similar capability inference diagram shown in the below figure.



The (piecewise-linear) function shown minimizes the sum of squared fitting error, over all concave functions. In this model, we use the free CVXOPT software package for concave optimization based on the Python programming language

Figure 8: Least-squares fit of concave function to given data

Comparing with the least-squares fit of the convex function in figure 7, it is obvious that when we change the dataset to be opposite and the constraints (11) to (13), we can also obtain the capacity inference diagram.

Therefore, in our project, the convex regression also fit the traffic flow theory with the concave curve shaped by given data. This is one important piece of evidence for us to apply convex regression to our project. In the next section, we will describe and apply the flow-density model of traffic flow theory with convex regression.

4 Application

This project aims at comparing the capacities inference under different speed limits with convex regression from the traffic data collected following the concave function. We need to fit the real data to the theoretical model to validate whether this model is effective for the traffic flow environment. To my knowledge, there is no related work using the nonparametric convex regression to solve the VSL problem with real data. Most works about convex regression are solving the demand (Varian, 1982), production (Varian, 1984; Allon et al., 2007) problems.

In section 3, we have already introduced and proved that how to apply convex regression in one concave case – traffic flow theory. In this section, we will apply convex regression to traffic data in detail. This section is divided into two subsections. In subsection 4.1, we will describe and explain the data was collected and preprocessed from the Finnish government agency, the Ministry of Transport and Communications' administrative sector. In subsection 4.2, we will apply the convex regression to estimate the capability inference diagram to fit the traffic dataset.

4.1 Data Collection and Preprocessing

For this project, we have to collect the raw data from the government agency. In order to keep the integrity and the quality of data, we will use one google extension tool Simple Mass Downloader to help collect raw data efficiently and effectively. After collecting raw data, we will use the python programming language to preprocess and resample the raw data.

The google extension tool is public and free, everyone who needs to collect raw data from the website could utilize it via its official introduction documents from below the link.

https://gelprec.github.io/quick_start_v2.html

And the python programming script for preprocessing the raw data is created originally, although some logic and calculations in details may be obtained from those python documents, answers from the website CSDN or google search. The readers could access the whole original programming script from the Appendix.

As for the steps on data collection and preprocessing described in the whole subsections, there are four main procedures I divided as shown in the below figure, Data Collection, Data Download, Data Structure Creation, and Data Preprocessing.

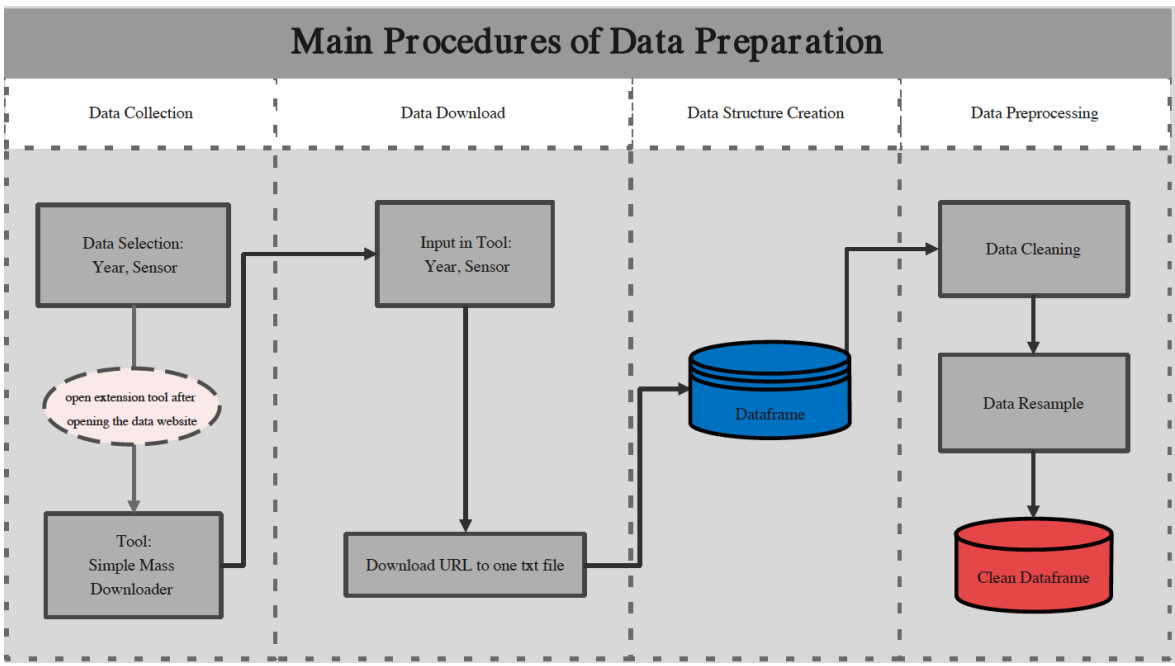


Figure 9: The main loop for data collection and preprocessing

For Data Collection, the users should first decide the year and sensor he needs and open the data webpage, then open the google extension tool Sample Mass Downloader and load page links; for Download, input ".csv" as Extensions, a file name such as "lamraw_116_20" as Text filter, then select all files and Export checked items to file; for Data Structure Creation, change text files to one dataframe divided by year and sensor; for Data Preprocessing, drop some unnecessary columns and null data, resample dataframe based on interval selected, after cleaning and resampling, now data could be analyzed.

4.1.1 Data Collection from Väylävirasto

This project collects the traffic data sets from the Väylävirasto which is the Finnish government agency that acts as the Ministry of Transport and Communications' administrative sector. The task of the Väylävirasto, Finnish Transport Infrastructure Agency (FTIA) is to be responsible for the implementation of the Finnish road and rail network and waterways. The agency was established at the beginning of 2010 and until the end of 2018 it was called the Finnish Transport Agency. The agency produces and publishes maps, statistics, publications, data sets, and open data related to traffic and fairways and all the materials are intended for use by all citizens and various actors in society.

This project aims at analyzing the estimation of the optimal speed limit to maximize the traffic flow at a given density, which is related to traffic data. The open data provided by the FTIA is the best validation of this thesis project. First of all, the open data almost covers the whole area of Finland: 01 Uusimaa; 02 Varsinais-Suomi; 03 Kaakkois-Suomi; 04 Pirkanmaa; 08 Pohjois-Savo; 10 Etelä-Pohjanmaa; 12 Pohjois-Pohjanmaa; 14 Lappi. In each area, there are lots of sensors collecting different transportation data to make the results of this project more convincing. In addition, the data sets are provided with high quality integrated by the FTIA, so that the data sets can ensure the accuracy, completeness, consistency, reliability of this project. What's more, the data sets last more than 20 years from 1995 to today and from the first day to the last day of each year, which can provide a high probability of comprehensive comparisons from time series.

As for the details of the data sets, the FTIA clusters them with years; in each year, the data sets are clustered into different areas; then for each cluster, all of the sensors' data are integrated from the first to the end day of this year. The data users can download the data no matter the day, the year, and the sensor. For example, if one user wants to download the transportation data of sensor 116 on the first day of the year 2019, he can first select 2019 and then select the area 01 Uusimaa then sensor 116 starting with 1, finally in the cluster 01 Uusimaa, find and download the CSV file from the URL address below.

https://aineistot.vayla.fi/lam/rawdata/2019/01/lamraw_116_19_1.csv

However, this project will collect large CSV files including target sensors and periods. The way mentioned above is not efficient, this project therefore will use one extension named Simple Mass Downloader which is one light batch download manager with flexible

filtering/mass-renaming options that make downloading files a lot easier & productive. In order to explain the data collection via Simple Mass Downloader in detail, an example of download setting is shown in Appendix B.

From the above step, we download all yearly data into one txt file, respectively. These text files include the data URLs of each day in this year as mentioned above. Because we need to store these data into one dataframe, we should download real data from this txt file with URLs and turn them to be a CSV file. One code example of downloading data from URLs file to CSV file has been shown in Appendix C and for other yearly data, the user could use the same code by changing the year and sensor.

4.1.2 Data preprocessing with Python

About the raw data, there are 16 features for each data point as shown below.

Table 1: The features of raw data

The Features of Raw Data	
❖ Sensor Identifier	❖ Lane of the Passing Car
❖ Year	❖ Direction (1 or 2)
❖ Day of Year (DOY)	❖ Vehicle Type
❖ Hour	❖ Speed (km/h)
❖ Minute	❖ Faculty
❖ Second	❖ Occupancy (presumed)
❖ Hundredth of a Second	❖ Interval
❖ Vehicle Length	❖ Distance to queue

First of all, for raw data, there is no name for each column, the user should set column name for dataframe. After that, for each CSV file, the size could be more than 1G which is large to process, causing too much time to process one file. In order to solve this case, the user needs to drop some unnecessary columns. In order to identify sensor and direction, the user should keep the column Sensor Identifier and Direction (1 or 2). Meanwhile, based on equation (5), the user should keep related time columns including Year, Day of Year (DOY), Hour, Minute as well as the Speed (km/h) to calculate the flow rate and density. As for the other time columns Second and Hundredth of a Second, on one side, in order to simplify the calculation of time, the user could drop both columns' data; on another side, there is no influence on the results that our project needed. In this project, we need to calculate the total numbers of passing cars in one certain time period (interval), it's inefficient to set the interval to seconds because there exist many data points with zero passing cars, causing the complexity of calculation. Therefore, the user only needs to remain the features shown in the table below.

Table 2: The remaininng features

The Remaining Features	
❖ Sensor Identifier ❖ Year ❖ Day of Year (DOY) ❖ Hour	❖ Minute ❖ Direction (1 or 2) ❖ Speed (km/h)

From the remaining features, it's obvious that there are four columns describing the time and even using the Day of Year to indicate the exact date in one year. For simplifying the calculation of the time period, the user should first change the Day of Year (DOY) to the exact date in one year and then combine the time in one column like "yyyy-MM-dd hh:mm:00".

As for how to transfer Day of Year (DOY) to the exact date in one year, because this project uses Python language to process the data, the user could introduce the TimeDelta structure directly from Python. A timedelta object represents a duration, the absolute difference between two dates or times, expressed in difference units (e.g. days, hours, minutes, seconds). For example, in the second day of the year 2019, the DOY is 2, we could get the exact date 2019-01-02 by using the function to_timedelta, which converts an argument from a recognized timedelta format / value into a Timedelta type.

By combining the second column 'Year' and third column 'Day of Year (DOY)' the user could get the exact date of this year like 'yyyy-MM-dd'. In this project, we need to analyze the function between flow rate and density, so the user also needs to integrate the 'Hour' and 'Minute' columns, now the user could obtain one column about time with 'yyyy-MM-dd hh:mm:00'. To calculate the flow rate in one certain period, this project will set the time into the Index column.

Considering that sometimes the different directions (1,2) have different flow rates because of, such as the rush hour, this project will divide the dataframe into two structures with direction 1 and 2. In this way, each dataframe has column Date Index, Sensor Identifier, Direction, and Speed.

Based on functions (5), we need to first decide the time interval. The time interval is the aggregation period which means based on this period, we could calculate the numbers of vehicles passing the measurement location. In this project, we choose 5 minutes as the time interval, which keeps the same estimation in Juho Ylimartimo's work in 2019.

As for any raw data, there must be some problem with meaningless data in the dataset. Usually, we use the interpolation method to replace them, such as zeros and non-numeric values, there is no doubt that interpolation is one important method to simulate the integrity and quality of data. Another way is to delete these meaningless data directly if these data don't have much influence on the results of the project and the interpolation method will simulate one result with a big error.

In this project, we use the second way. First of all, the data we collect is about the vehicle information only including two measures dimensions, time and speed, about the time, the errors of interpolating one time between two data and deleting the data directly will keep similar. The important part is the speed, when using an interpolation method, such as one-degree polynomial interpolation which will simulate one median speed between two data, the user would create a bigger error. Considering the real situation, the speed of one vehicle between the past one vehicle and the coming one is not always the median value between the two speeds. Therefore, simulating the speed of one vehicle is not a good way in this project. Besides, in this project, we have tested that there is no too much meaningless data in our raw data, for each dataset, the meaningless data only account for around 0.2%. Therefore, there is no need to use the interpolation method in this project.

After cleaning the raw data, we could obtain two dataframe for each yearly information of each sensor as shown below.

Table 3: Two DataFrames

Two DataFrames	
❖ Date (Index) ❖ Sensor Identifier ❖ Direction 1 ❖ Speed (km/h)	❖ Date (Index) ❖ Sensor Identifier ❖ Direction 2 ❖ Speed (km/h)

Based on the equations (2)-(5), we need to calculate the flow rate and density, however, the table we have shown above doesn't include both columns, so in the next session, we will describe how to calculate the flow rate and density based on the data we preprocessed in this session.

4.1.3 Calculation of Flow Rate and Density

From the data we calculated, it's obvious that we now only have the speed. In order to calculate the flow rate and density, we need to decide which method we need to use and to calculate the average speed.

The traffic data collected including the time and speed of the individual vehicle. After the user decide the aggregation period (time interval), we can compute the average speed and flow rate within the time interval directly via loops at a certain location (Victor L. Knoop & Winnie Daamen, 2017). Suppose we have I vehicles passing the measurement location with a speed v_i, then the flow rate is vehicle number I, and the arithmetic mean of the speeds (also called time mean speeds (TMS)) could be computed with the following formula:

$$\upsilon = \frac{\sum_i v_i}{I} \tag{14}$$

However, TMS is only exact when computing the homogeneous speed of passing vehicles, which means that in one road the speed of all vehicles should keep almost even which is impossible in real data. Using this method (TMS) will cause a bias in the speeds (time mean speeds are too high), and a bias in the density (the computed densities are too low), especially for higher densities (Knoop, Hoogendoorn, and Van Zuylen 2007) for an explanation and a comparison of real data.

In order to improve the bias, Edie's generalized definitions can be used (Edie 1965). Generally, they introduced how to collect observations in an area in space-time into a flow, density, and speed. They integrated their calculation like the below method. As we described at the first session, the total distance D traveled by all vehicles and the total time spent in the area T is calculated; the area of observation has a size of A (in units of space-time). The flow and density then are calculated by q = D/A and k = T/A. The speed is defined as v = q/k.

Because we only have speeds of these passing vehicles, we now need to transfer these functions with only speed. Suppose we have I vehicles passing the measurement location with a speed v_i and let us consider an area in space-time, with length λ and time interval τ. The area is A = λτ. The total distance covered is D = $\sum_i \lambda = I\lambda$; the total time spent is T = $\sum_i \lambda / vi$. Now, the speed can be computed with the below equation, the inverse of the mean of the inverse of the speeds of individual vehicles.

$$v = \frac{q}{k} = \frac{D/A}{T/A} = \frac{I\lambda}{\lambda \sum_i 1/v_i} = \frac{I}{\sum_i 1/v_i} = \frac{1}{\overline{1/v_i}} \tag{15}$$

This is the famous space mean speed (SMS) that is roughly equivalent to the harmonic mean of those (Knoop & Daamen, 2017; Maerivoet & De Moor, 2005), SMS is the total distance traveled divided by the average travel time of the individual vehicles.

However, although we have already calculated the space mean speed and flow rate from the raw data, we still cannot calculate the density directly from these two variables without considering any assumption. As mentioned in Maerivoet & De Moor, 2005, there exists a unique relation between three of the previously discussed macroscopic traffic flow characteristics density k, flow rate q, and space-mean speed v:

$$k = \frac{q}{v} \tag{16}$$

This relation is also called the fundamental relation of traffic flow theory, as it provides a close bond between the three quantities: knowing two of them allows us to calculate the third one (note that the time-mean speed in equation (14) does not obey this relation). In general, however, there are two restrictions, i.e., the relation is only valid for (1) continuous variables, or smooth approximations of them, and (2) traffic composed of substreams (e.g., slow and fast vehicles) that comply with the following two assumptions:

(1) **Homogeneous traffic**: The relation is valid for continuous variables or smooth approximations of those. (Maerivoet & De Moor, 2005).
(2) **Stationary traffic**: All the substreams in the observed traffic need to be homogenous and stationary. I.e., the substreams comprise only of same-typed vehicles and the vehicle trajectories remain constant over time and space (Maerivoet & De Moor, 2005).

As for the two assumptions, Juho Ylimartimo (2019) also explained that using the resampled flow rates and average speeds from a sensor just 500 meters away from a bottleneck (such as light-controlled intersections and junctions) to calculate the vehicle density over a kilometer distance most likely results in a skewed estimate of the local traffic state. Actually, in this project the data collected is on the high-speed road, there are no too

many or even no light-controlled intersections and junctions. Although there exist some sub-streams or some traffic accidents causing occasional junction which have some influence on the traffic condition, in general, the traffic keeps homogeneous and stationary.

After considering the two assumptions, we can calculate the density from the formula (16). However, it's obvious that in this function the speed is the denominator, which could not be zero. But for real raw data, for example in the midnight of one day, there exists no vehicle passing the measurement location in one specific time interval, so there is no doubt that space means speed is zero. As we all know that in the python model the calculation would appear zero-division errors when the raw data includes null/zero data. For null data, we have deleted directly based on the explanation above in the data preprocessing session. For zero data of flow rate when calculating the density, we have to pay attention to the definition of space mean speed in creating models. In the model, we need to separate them into two situations, which is for zero flow rates in one specific time interval, the other one is for normal flow rate situation.

After data preprocessing and calculation of flow rate and density, the final two DataFrames are composed of five columns like below tables shown.

Table 4: Final Two DataFrames

Final Two DataFrames	
❖ Date (Index)	❖ Date (Index)
❖ Sensor Identifier	❖ Sensor Identifier
❖ Direction 1	❖ Direction 2
❖ Flow Rate (estimated in 5 minutes)	❖ Flow Rate (estimated in 5 minutes)
❖ Density	❖ Density

4.2 Estimation of the Optimal Speed Limit to Maximize the Traffic Capacity at Given Density with Convex Regression

This project infers the different traffic capacities under different speed limits with the processed traffic data flow rate and density shown in table 4 above with the application of convex regression. However, the shape of the flow-density function follows the concave structure. In this thesis, the method of convex regression in the concave case is used to infer theoretical parameters relating to highway performance: the capacity q and the critical density k.

In subsection 4.2.1, this project will select several sensors and illustrate the flow-density function with yearly and monthly data; In subsection 4.2.2, this project will explain the utilization of convex regression in traffic flow theory in detail including the theoretical algorithm and programming script.

4.2.1 Flow-density Illustrations with Real Data

This project aims at estimating the optimal speed limit to maximize the traffic capacity. Therefore, selecting yearly and monthly datasets of sensors and illustrating them would be absolutely presentational to readers to find the differences among different speed limits.

The sensor location information could be obtained from the combination of an interactive map hosted by Eero Salminen[1] and the books kept by Väylä[2]. Considering the balance between the credibility of data results and the complexity of description in this thesis, this project selects four sensors from several different highways with variable speed limits in one direction and illustrates the yearly datasets.

As for the selection of sensors in different highways is to obtain different speed limits dataset. Normally, in one highway, the speed limit keeps the same for most areas. In different highways, it's obvious to obtain different traffic capacities under different speed limits in vehicles. Of course, the programming script fits all sensors and the users can validate the results shown in this project with other datasets by selecting several sensors with variable speed limits.

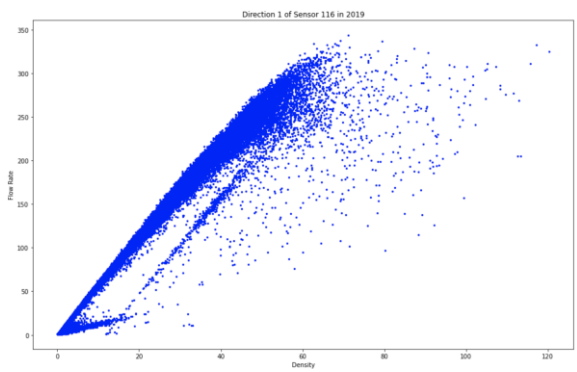
The information and reasons selected are shown in the below table. As for one of the reasons for selecting sensors, we will avoid these sensors close to onramp and offramp where always happen accidents causing significant influence on the traffic flow and density of results. In addition, this project aims at inferring the differences of traffic capacities at different speed limits, so the speed limit is considered when selecting sensors to avoid the homogeneity of the speed limit, we select four sensors with different speed limits of 70km/h, 80km/h, 90km/h and 100km/h. We collect these from the official book mentioned above, and we also validate the information by driving a car and passing the sensor to collect the speed limits and distances from both on-ramp upstream or off-ramp downstream. More information will be shown in the below table.

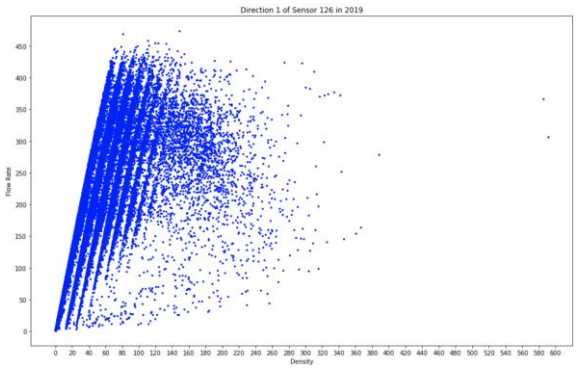
1. https://www.arcgis.com/home/webmap/viewer.html?webmap=10d97c7d9d9b41c180e6eb7e26f75be7
2. https://vayla.fi/tilastot/tietilastot/lam-kirjat#.XZsbdCVS9Z1

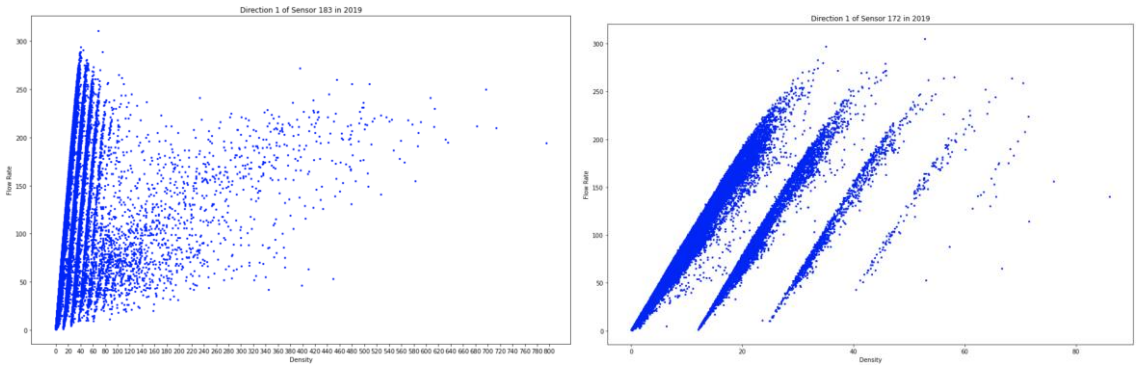
Table 5: Four Sensors Sellected

Sensors Information and Reasons of Selection	
❖ Sensor 116 ➢ Direction 1: Itäkeskus – Tapiola ➢ Location: Leppävaara, Espoo ➢ Characteristic: speed limit 70km/h and at a few hundred meters distance from both an on-ramp and an off-ramp	❖ Sensor 126 ➢ Direction 1: Itäkeskus – Tapiola ➢ Location: Konala, Helsinki ➢ Characteristic: speed limit 80km/h and within a kilometer distance from both an on-ramp upstream and an off-ramp downstream
❖ Sensor 183 ➢ Direction 1: Vantaa – Espoo ➢ Location: Järvenperä 2, Espoo ➢ Characteristic: speed limit 90 km/h and no on-ramp upstream and off-ramp downstream within a kilometer distance	❖ Sensor 172 ➢ Direction 1: Porvoo – Helsinki ➢ Location: Helsinki ➢ Characteristic: speed limit 100km/h and no on-ramp upstream and off-ramp downstream within a kilometer distance

Based on the resampled datasets, it's easy to plot the flow-density scatterplots of these sensors selected in table 6 above. Normally, it's better to observe the different results in one figure of different datasets. However, considering the big size of each dataset and the situation that theses plots are scatterplots which would cause big overlapping area if plotting all datasets in one figure, so this project plots these scatterplots in four separate figures (a-d), so there exist different ticks in each figure for different sensors. However, that method will not bring too much difficulty in observing the results.







Above four figures are the flow–density scatterplots of four sensors selected at direction 1. The density is the density in per kilometer and the flow rate is estimated in 5 minutes time interval. The upper left is the scatterplot of sensor 116 with speed limit 70 km/h, the upper right is the scatterplot of sensor 126 with speed limit 80 km/h, the bottom left is the scatterplot of sensor 183 with speed limit 90 km/h and the bottom right is the scatterplot of sensor 172 with speed limit 100 km/h .

Figure 10: the yearly flow-density scatterplots of four sensors

4.2.2 Application of Convex Regression in Traffic Flow Theory

The data collected is a concave dataset in this project. In order to fit a convex function to given data, we need to process the data before using the CVXOPT optimization library. In subsection 4.1, we preprocess the raw data into a yearly unit. Considering the computing storage of CPU and time cost of obtaining the results, we need to divide the yearly data into monthly data first which is enough to compute with local CPU or any website programming tool.

After we divide yearly datasets into small monthly datasets. We need to order the data by ascending densities. Although we have monthly data now, the dataset still keeps huge, because the size of the matrix would be too sensitive on the number of the data, so we need to control the size of the data. In addition, when we use the solver model to solve the optimization problem, it's a good way to aggregate these data linearly dependent first. Therefore, in this project, we need to remove these data that are linearly dependent.

However, because the densities keep decimal format causing it difficult to do data aggregation. Therefore, firstly we need to round the densities to integers, in this way, there will appear the same densities then group by these densities, we calculate the mean of the flow rate. After that, we both keep the integrity of data and avoid the high time cost in running the model.

Because we use the programming script as an example provided by Boyd & Vandenberghe in their book to fit the concave function to given data. Therefore, the first step is to change the objective function and the constraint to fit the concave function at the given data.

In this project, the objective function is the equation (10) and the restrictions should be the inequality constraint (13). In order to fit the objective function and inequality constraint to the standard form of quadratic programming so that we can adjust the matrix coefficients of variables. In order to make it more readable and easier to understand the whole model in concave function, I use the same notations of variables shown in the programming script provided by Boyd & Vandenberghe. We set the variables yhat(m), g(m), the objective function is

$$(yhat - y)^2 = \frac{1}{2} * 2 * yhat^2 + (-2y) * yhat + y^2 \tag{17}$$

So, the standard objective function should be like:

$$minimize\ \frac{1}{2} * 2 * yhat^2 + (-2y) * yhat + y^2 \tag{18}$$

Because the collected data follow traffic flow theory and concave function, we need change the shape of the inequality constraint like below:

$$subject\ to\ \ \hat{y}_j \leq \hat{y}_i + g^T(u_j - u_i), i,j = 1,\dots,m \tag{19}$$

According to the standard form of constraint in QP, the constraint in this project should be formed like below:

$$\begin{gathered} -yhat[i] + (u[i] - u[j]) * g[i]' + yhat[j] \leq 0 \\ j, i = 0,\dots,m-1. \end{gathered} \tag{20}$$

Combining the objective of Boyd & Vandenberghe and this project, we need to change the coefficient of matrix P to 2 by multiplying 2 with the original matrix. Meanwhile, it's obvious that the coefficient of one-dimensional variable yhat (m) is -2, so we need to multiply -2 with the original vector. Note that the original matrix and vector are referred to these in the model of Boyd & Vandenberghe in 2004. These are the changes in the objective function of fitting this project.

From the constraint (20), we can notice that the coefficients of yhat (m) and g (m) are changed from the model used by Boyd & Vandenberghe in 2004. Fortunately, it is easy to know how to change the coefficient of these variables because the whole inequality constraint only changes the shape.

Now, this model fit the concave function to the given data. We can use this model to fit the data and obtain the capacity of inference with convex regression. Figure 11 shows the monthly scatterplots and the corresponding capacity inferences with convex regression in October 2019 for each sensor we selected.

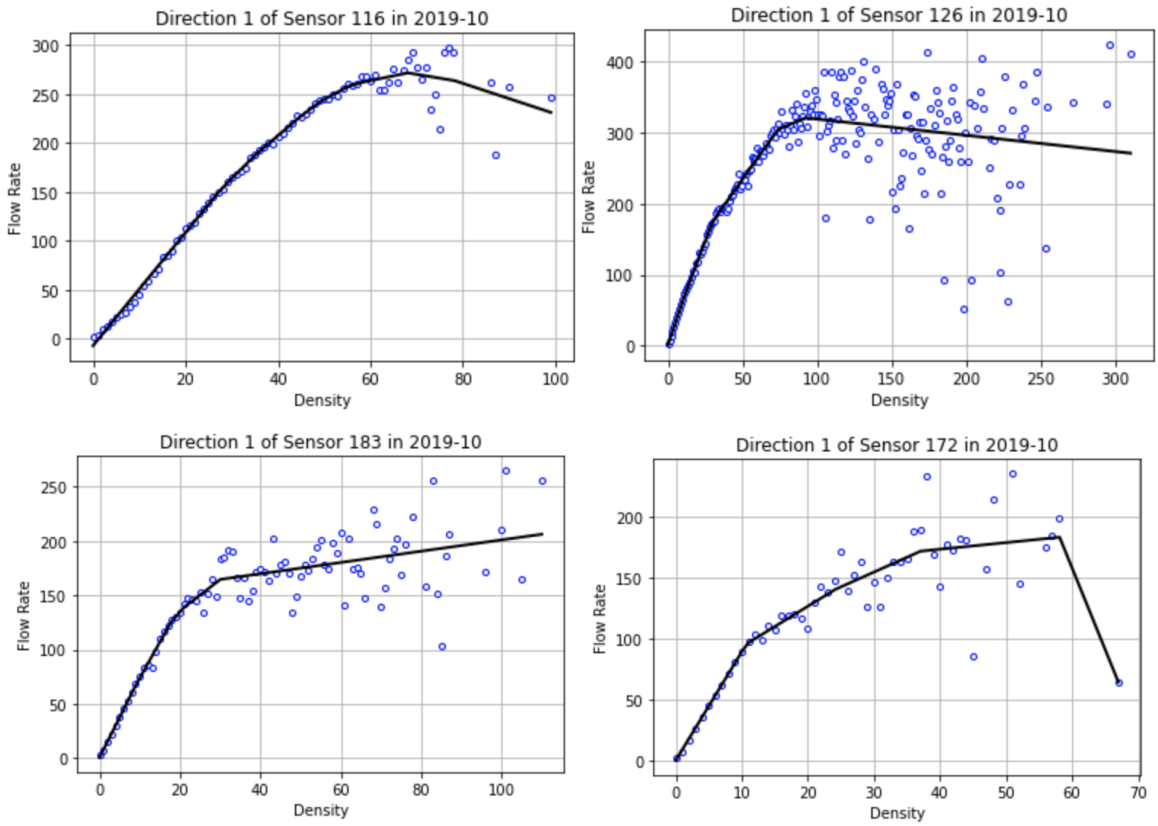


Figure 11: the capacity inference diagram of four sensors in October 2019

5 Results and Discussion

This section will describe the results of the application of traffic data with transformed convex regression from section 4. In section 4, we obtain the capacity inference diagrams for each sensor under variable limit speeds. In order to find the differences of flow-density models under variable speed limits, we need to combine these flow-density models in one figure to estimate the cross points.

This section will be divided into three parts. In the first subsection 5.1, we will provide a general statistical and summary of the experimental data. In subsection 5.2, we will analyze the monthly flow-density models of four selected sensors, respectively, which is related to the first research question:

(1) Does the flow-density model keep similar critical capacity varying in the monthly change? If no, what's the reasons for these differences?

In subsection 5.3, we will make comparisons on flow-density models among four sensors to estimate the cross points, which is related to the second research question:

(2) Does the flow-density model keep different critical capacities varying in the speed limits? Is it possible to estimate the optimal speed limit to maximize the traffic flow at given density?

5.1 Statistical Analysis of Experimental Data

In this project, we select four sensors with variable speed limits from 70 km/h to 100 km/h and we divide the yearly dataset into 12 partial monthly datasets for each sensor respectively. In this subsection, we will discuss the datasets from two parts. The first part will describe the monthly flow rate of four sensors. Under this statistical analysis, we will know the general traffic capacity for each sensor. The results are shown in the below table.

Table 6: Monthly flow rate and median slope of four sensors

Monthly Flow rate	Sensor 116 (70 km/h)		Sensor 126 (80 km/h)		Sensor 183 (90 km/h)		Sensor 172 (100 km/h)	
Median Slope	Flow rate	Slope (Median)	Flow rate	Slope (Median)	Flow rate	Slope (Median)	Flow rate	Slope (Median)
January	857,914	[5.25,5.78]	1,131,397	[6.00,6.72]	601,249	[6.93,7.76]	431,114	[7.18,8.32]
February	812,506	[5.31,5.85]	1,049,130	[6.30,7.07]	553,807	[6.75,7.55]	400,968	[7.33,8.54]
March	934,675	[5.04,5.78]	1,202,161	[6.34,7.09]	633,687	[6.85,7.67]	460,774	[7.98,9.02]
April	923,573	[5.39,6.00]	1,180,992	[6.11,6.84]	646,356	[6.85,7.67]	477,051	[8.26,9.40]
May	976,232	[5.58,6.31]	1,242,449	[6.52,7.32]	703,514	[7.58,8.41]	530,527	[8.21,9.51]
June	907,094	[5.10,5.95]	1,160,043	[6.07,6.77]	668,103	[7.38,8.27]	527,477	[7.95,9.05]
July	38,496	[2.44,4.00]	1,061,726	[5.97,6.70]	642,309	[6.72,7.55]	523,992	[8.56,9.73]
August	10,426	[2.42,3.99]	1,236,470	[6.15,6.90]	706,867	[6.90,7.74]	561,068	[8.35,9.48]
September	10,989	[2.43,4.01]	1,236,907	[6.19,6.93]	684,860	[7.17,8.04]	512,601	[8.92,10.17]
October	470,174	[4.75,6.26]	1,279,996	[5.84,6.79]	704,047	[7.02.7.88]	520,469	[8.15,9.26]
November	931,212	[5.43,6.14]	1,214,064	[6.07,6.78]	653,907	[6.13.7.62]	477,728	[7.40,8.59]
December	787,481	[5.29,6.03]	1,084,904	[6.03,6.77]	555,793	[6.88.7.73]	436,946	[7.81,8.90]
Total	**7,660,772**	**[4.78,6.35]**	**14,080,239**	**[5.84,6.80]**	**7,754,499**	**[6.08,7.58]**	**5,860,715**	**[8.08,9.42]**

The traffic data of this project is collected from the highways. The total monthly flow rate of each sensor will be an important indicator when analyzing the experimental data. The total flow rate scope of the sensor can be influenced by the population size of the adjacent city or area, if the population size is big the monthly flow rate will be big.

In addition, the destination of the route has a great impact on the monthly flow rate scope. For example, the route to the city center such as shopping centers, supermarkets and entertainment places or working centers such as institutions and some other similar places where many people work. Under these situations, the monthly flow rate will be higher than other sensors. Sometimes, it is difficult for people to split these two variables (the population size and destination) and even other possible variables because the higher population size area and busy destination could be overlapping.

In this project, the monthly flow rate of sensor 126 always keeps higher than any other sensor from January to December. The sensor 126 is located in Konala, Espoo with Tapiola destination. The Adjacent area of Tapiola is one place with shopping centers such as Stockmann, institutions such as Aalto University. And sensor 126 is located in ring I of Finland transportation, which means this route is the busiest comparing with other routes.

Only when the flow rate keeps higher, the congestion will happen to this area. Therefore, based on the statistical results of sensor 126, we could obtain the critical capability from the dataset.

For sensor 116, which has the same destination and is also located in ring I but between sensor 116 and 126 there is one off-ramp which directs to the big shopping center Sello, Espoo. So, it is obvious that the monthly flow rate always keeps lower than sensor 126. And even in July, August, and September in 2019, the monthly flow rate keeps extremely low than the normal monthly flow rate, which belongs to the abnormal dataset that we will not consider these monthly flow rates in the next subsection. There is no evidence that explains the reason for this abnormal situation, it could be road maintenance.

Considering the same variables (the population size and destination) for sensor 183 and 172, it is reasonable that the monthly flow rate keeps lower than sensor 126 and even sensor 116 except for these three abnormal months, especially the sensor 172, the yearly flow rate only has 5,860,715 because it is located in the area far away from the city center and some busy destinations.

After analyzing the possible variables of the monthly flow rate from population size and destination, there is also one important indicator which has a great impact on the monthly flow rate, and it is the season and the public holidays, but in this project, we don't analyze the flow rate under the public holidays in Finland. In different seasons people have different willing to travel. For example, in winter people would like to stay at home instead of traveling outside and in spring and summer the opposite situation. Therefore, the monthly flow rate in spring and summer should be higher than the winter. And the experimental data shown above also reflects this seasonal difference.

We just analyze the possible variables from the experimental data and validate some reasonable points with the statistical results. In this project, we need to consider the variable speed limits for each sensor. The slope data shown in the table mean the ration of flow rate and density in the flow-density diagram. We divide the ration of each sensor in each month into 10 bins and select the median ration (slope) here to estimate the general shape of the flow-density diagram. The speed limits have an obvious impact on the median slope of the flow-density diagram. The higher the speed limits, the higher the median slope scope. We can take the yearly data on the table above as an example. The slope scope of sensor 116 with speed limit 70 km/h keeps between 4.78 and 6.35, the slope scope of sensor 126 with speed limit 80 km/h keeps between 5.84 and 6.80, the slope scope of sensor 183 with speed limit 90 km/h keeps between 6.08 and 7.58, the slope scope of sensor 172 with speed limit

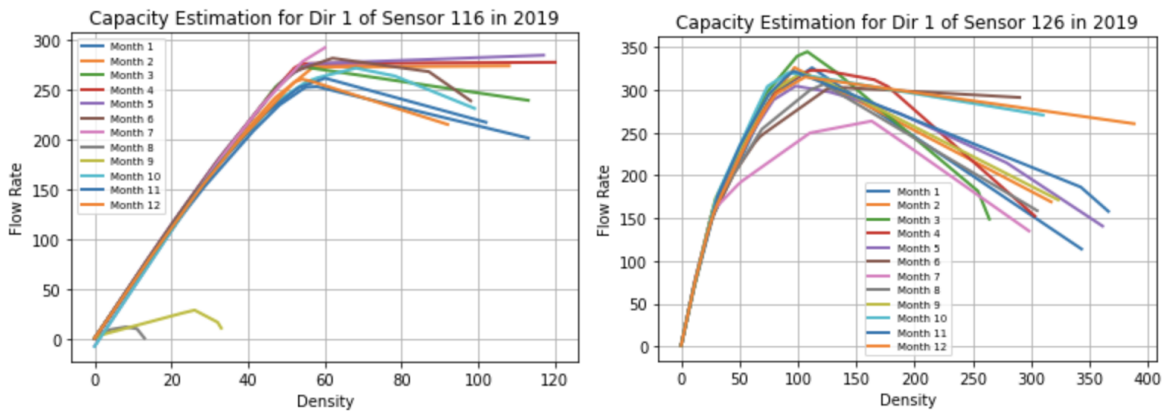
100 km/h keeps between 8.08 and 9.42. The low bound and high bound of each sensor keep increasing of these four sensors. By combining the yearly flow-density scatterplots of four sensors shown in figure 11 and statistical results, we can generally estimate that the sensor with the highest speed limit will approach the critical capacity steeply when controlling the yearly or monthly flow rate.

Finally, when we analyze the experimental data with both yearly flow rate and variable speed limit, there is one significant situation we need to consider. Whether the flow rate can approach to the critical capacity under this speed limit. It is not difficult to understand the situation that when the speed limit is very high and the monthly flow rate keeps low, the free flow speed which means the speed to keep the highway smooth and no congestion will be lower than the speed limit, which causes that there would be no critical capacity because there is no congestion in this area. Therefore, we need more analysis on this issue from the next subsections.

5.2 Monthly Comparisons of Flow-density Model

Although this project considers the flow-density models in variable speed limits and we should compare the model among different sensors under VSLs. But only consider this way will miss the other perspective that under the same speed limit, is there any significant change of the flow-density model? Therefore, the monthly comparisons of the flow-density model for each sensor themselves also are important.

After using the CVXOPT optimization algorithm, we can get the capacity inference for direction 1 of each sensor this project selected in the year 2019, respectively. As mentioned above, because this project aims at finding the differences between the flow-density models in variable speed limits with real traffic data. Therefore, we need to make comparisons to obtain our results. In this project, we will compare them from two parts. In one part, we compare the monthly flow-density model for each sensor, respectively. In this way, we can know the differences of flow-density model among different months for this sensor in one year. If we obtain the significantly different shape of flow-density models for each month in one year, we need to focus on the reasons for the differences. Because there are different findings among these four sensors, we then cluster four sensors into two parts for analysis.



Above two figures are the flow–density models of sensors selected at direction 1 with variable speed limits: the sensor 116 with speed limit 70km/h, the sensor 126 with speed limit 80 km/h. Each sensor with whole yearly traffic data is divided into 12 months and each line illustrates the capacity inference of flow-density model in this year.

Figure 12: the capacity inference of flow-density model for sensor 116, 126

For both sensors 116 and 126, the flow rate first keeps increasing and then decreasing when it approaches the critical point. And combing the statistical analysis above, we can conclude that there exists a congestion region for sensors 116 and 126.

Before the congestion condition, for sensor 116, it's obvious that the monthly flow-density model almost keeps the same curve except for these months July, August, and September. Based on the statistical results we discussed in subsection 5.1, the monthly datasets of these three months of sensor 116 are abnormal which cannot infer the potentials correctly, we need to ignore these monthly datasets. For the critical capacity of monthly flow-density models, the flow rate is almost practically the same throughout the year between 250 and 270. Interestingly, even the monthly critical density stays at 60 for the whole year.

Similarly, for sensor 126, the monthly critical capacity mostly keeps a similar scope between 300 and 350 among twelve months. In addition, it is obvious that monthly critical density also mostly stays at one certain point 110.

However, the slopes of the congestion region above the critical density reveal large monthly differences for both sensors 116 and 126. The slope also means space mean speed in the flow-density model. From both sub-figures shown above based on the convex regression, it's obvious that in those months with good weather conditions the slopes mostly keep gentler than bad weather after critical density. It is quite intuitive that in summer or autumn months with better weather conditions and more light the flow rate in the congestion region is much higher than during the winter months when there are rain and snow and dark throughout the day.

In order to compare the seasonal differences clearly, we also draw the seasonal flow-density scatterplots for sensor 116 and 126. For winter datasets, we choose months of December, January, and February for both sensors with blue points shown in figure 13 below. For a season with good weather conditions, we select months of April, May, and June for sensor 116 because of the abnormal dataset in summer and months of June, July and August for sensor 126 with red points shown below. We can conclude the same inference with the results based on convex regression that the flow rate above the congestion region is much higher in good weather condition than bad.

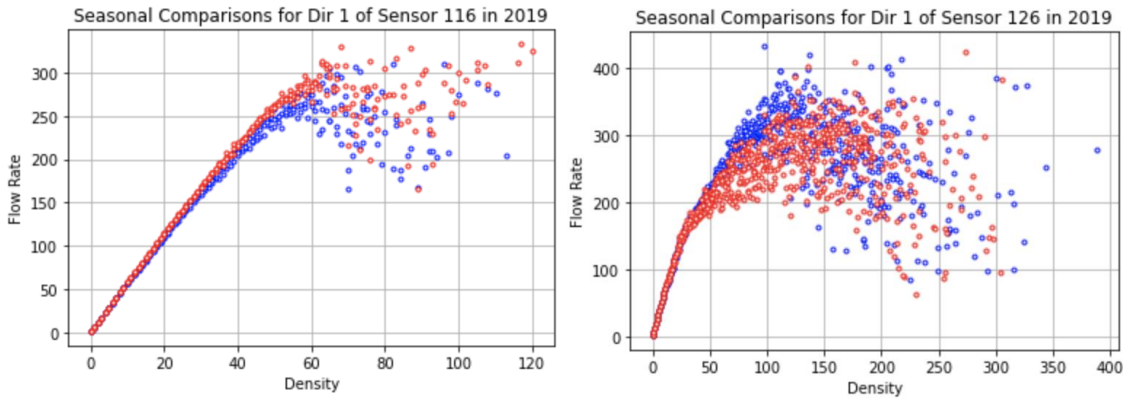


Figure 13: the seasonal flow-density scatterplots for sensor 116, 126

The two sensors 116 and 126 analyzed are these sensors with critical capacity under the speed limit. However, for sensor 183 with a speed limit of 90 km/h and sensor 172 with the speed limit of 100 km/h, it's possible that there is no critical capacity because of the high-speed limit and low flow rate.

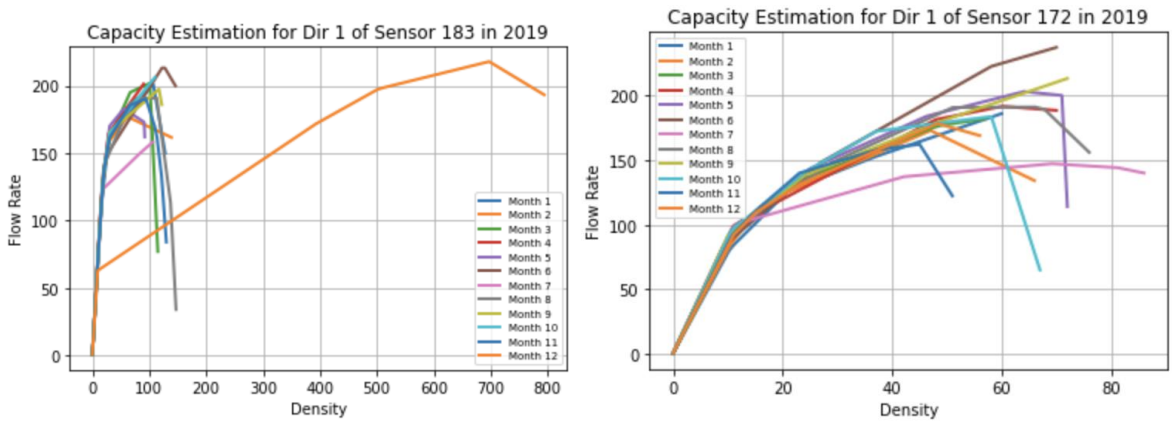


Figure 14: the capacity inference of flow-density model for sensor 183, 172

For sensor 183, obviously, there is one abnormal monthly flow-density model in December. We need to analyze the raw data to find the possible reason for this abnormal distribution of the flow-density model. We select these raw data with a density higher than 300 in December and find that all of these data are located in the period from 6:30 am to 8:30 am during working days which is the busy time for traffic because people drive to work during this period. And these data only appear during working days that validate the suppose. Therefore, there exists congestion in this time period during the working days in December. However, this kind of situation only appears in the month of December in the whole year, which means there are some changes for the road station, such as the road maintenance

causing the reduction of the lane, because in normal situation there exist two or three lanes in one direction for cars to pass. The special cases such as road maintenance will reduce the number of lanes in this road, so under the similar monthly flow rate shown in table 6 but with the significantly different monthly flow-density model, the reason for this kind of congestion cannot be considered for our analysis. Therefore, we should ignore the monthly flow-density model in December of sensor 183.

For sensor 172, there are two different clusters shown from the model above for first sight. One cluster indicates the flow rate approaches the critical capacity because, in some models, it appears to decreased capacity when the flow rate approaches one highest point, such as the month May, October. After checking with the raw dataset shown in Appendix D, we find that there exists only one abnormal data point causing the direction change of the whole flow-density model. Under this situation, we need to remove the abnormal data point.

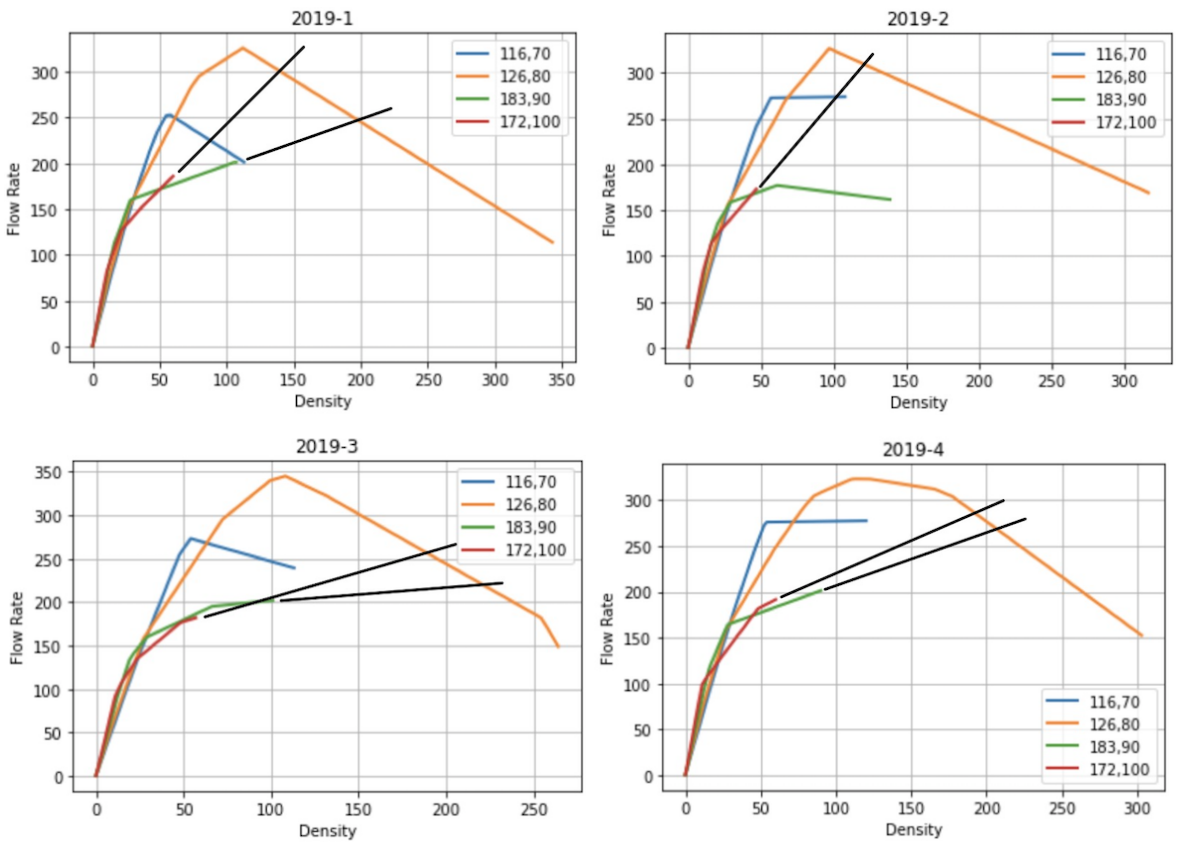
After removing the abnormal data points we can notice that in sensor 172 with a speed limit of 100 km/h, the monthly flow rate doesn't approach the critical capacity because the flow rate still keep increasing when the density keeps increasing, which means that in normal traffic condition, there is no traffic congestion situation and critical capacity for sensor 172 with speed limit 100 km/h, which means that the speed limit setting here is meaningless.

In summary, for the first research question, we can conclude that below the critical density, the flow rate is practically the same throughout the year, and the critical density (or maximum capacity) is also very stable across months. However, the slopes of the congestion region above the critical density reveal large seasonal differences. It is quite intuitive that in summer months with better weather conditions and more light the flow rate in the congestion region is much higher than during the winter months when there are rain and snow and dark throughout the day.

5.3 Monthly Comparisons of Flow-density Models under VSLs

In subsection 5.2, we conclude that the slopes of the congestion region above the critical density reveal large seasonal differences. Therefore, after we compare the monthly flow-density models in one year for each sensor, we then need to compare the flow-density model between different sensors with different speed limits, which is the most important part to know whether we can estimate the optimal speed limit from the flow-density models.

In this part, we cluster these datasets of four sensors and illustrate them in one plot to compare them clearly and we can estimate the cross points. Considering the possible different traffic flow in different months because there are many variables we need to control, such as seasonal weather we mentioned in 5.2 or public holidays in one year, for these four sensors we compare them in the same month. In other words, we compare the monthly flow-density models of four sensors by months. In order to avoid the duplicate analysis, we just select four groups from January to April to analyze the results shown in the below figure 15. Notice that we remove the abnormal data points mentioned in 5.2 and estimate the general flow-density diagram for sensor 172 and 183, we choose the same slopes with the previous stage to extend the diagram until they have cross points with sensor 126.



Above four figures are the flow–density models of four sensors selected. The legend of each plot means the sensor identifier and the speed limit. The black lines are corresponding estimation of flow-density diagram.

Figure 15: Comparisons of flow-density model for four sensors

From the discussion in subsection 5.1 and 5.2, we have already known that the flow rate approaches the critical capacity for sensor 116, 126. Obviously, from the four flow-density models we can notice one important point when just comparing the sensor 116 and 126. At the first stage, the density scope is between 0 and around 40, we can see that the slope of two curves keep the same distribution, which means there is no big difference when the flow rate stays in one low level although there are variable speed limits for different sensors.

After the first stage, the slope of the two sensors begins to change. Sensor 116 with a speed limit of 70 km/h becomes more sensitive before it approaches its critical capacity and keeps a high flow rate with the same density. We can describe the theory from one sub-figure in detail. Taking the first sub-figure shown above, we can see that the cross point of sensor 116 and 126 is around（63, 259）, based on the flow-density model under VSLs mentioned in subsection 2.1, when the traffic density is lower than the cross point 63, we can just set the speed limit with 70 km/h instead of 80 km/h. In this arrangement of speed limit setting, we can reduce the possibility of accidents by lowering the driving speed of drivers. For the other three sub-figures shown above keep the same property, at lower density we can reduce the speed limit.

After sensor 116 approaches the critical capacity at around the highest point (55, 250) and the flow rate begins to decrease when the density keeps increasing. Because of the limit of the raw data and the method transformed convex regression, we cannot estimate and draw out the right part of sensor 116, but it doesn't have a big influence on the analysis in this project. Because after the cross points in the four sub-figures, the flow rate of sensor 116 with a speed limit of 70 km/h keeps lower than the flow rate of sensor 126 with the speed limit of 80 km/h. Therefore, we can know that the congestion of sensor 116 after density approaches to more than 63 appears, we need to change the speed limit to a higher one to reduce the impact on the traffic jam, such as the air pollution and expensive delay for people's special business work.

As for the monthly flow-density models shown above of sensor 183 with the speed limit of 90 km/h and sensor 172 with the speed limit of 100 km/h, due to the limit of raw data and the methodology applied in this project, we cannot draw out the right part of the flow-density model directly. In addition, the flow rate of sensor 183 and 172 do not approach their own critical capacity after we remove the abnormal data points in raw datasets, the flow rate still keeps increasing when the density keeps increasing. However, from the black line shown in figure 15, we can find that if we estimate the rest part of sensor 183 and 172 based on the part shown already with raw data, there exist some cross points we can use to decide the optimal speed limit at the given density like what we analyzed between sensor 116 and 126 with variable speed limits.

Actually, when comparing the four flow-density diagrams, we can notice that at the stage below around 60 density, the traffic flow would be the highest when the speed limit is 70 km/h, the lowest speed limit; At the stage where the density keeps in between around 60 and near to the critical capacity of sensor 126, the traffic flow would be the highest when the speed limit is 80 km/h; for the rest stages between 172 and 183, more investigation and simulation should be made to obtain the results. But from the earlier works, there should exist one certain scope of density to obtain the highest traffic flow.

Therefore, for research question II, the flow-density model keeps different critical capacities varying in the speed limits and it is possible to estimate the cross-points for different flow-density models with variable speed limits and estimate the optimal speed limit at the given density.

6 Conclusions

This project applies the convex regression to solve the traffic flow theory problem. We collect traffic data from the highway in Finland and estimate the optimal speed limit to maximize the traffic flow at a given density based on the real data using the convex regression. From the results of this project, we conclude that below the critical density, the flow rate is practically the same throughout the year, and the critical density (or maximum capacity) is also very stable across months. However, the slopes of the congestion region above the critical density reveal large seasonal differences. It is quite intuitive that in summer months with better weather conditions and more light the flow rate in the congestion region is much higher than during the winter months when there are rain and snow and dark throughout the day. In addition, it is possible to estimate the cross-points for different flow-density models with variable speed limits. The first cross scope is located near the critical density of the lowest speed limit, the second cross scope is located in near the critical density of the second lower speed limit. The third and fourth cross scopes are estimated in the critical density of sensor 183 and 172 based on Carlson's estimation. We can apply these results to improve traffic condition.

Although we obtain remarkable results from this project, there still exist some limitations for this project. We only collect the traffic data from highways which is different from the city roads with the traffic light. On highways, there is no traffic light which has a great impact on the estimation of the optimal speed limit. In addition, we collect historical data, for some abnormal conditions appeared in this raw data, there is no documents to explain the reasons in detail, so in this project, we just figure out the abnormal monthly datasets and ignore them while in order to analyze these situations in comprehensive perspective, we need to know the specific reason for these abnormal conditions. The third limitation is about the speed limits for the area of each sensor, we record the speed limits practically and this project lasts six months, we record speed limits two times, one is in September and the other in November, the speed limits keep the same. However, there is no document recording whether there exist different speed limits in different seasons in historical data. Therefore, the speed limits keep the same is one assumption in this project. Finally, this project just makes the simplest simulation for these datasets without a complete flow-density model from raw data. There exists a much more exact simulation method to estimate the whole flow-density models for each sensor, but this project does not focus on

the simulation method, then the simulation for the flow-density model is not considered, we just estimate the general flow-density model based on the shape of other sensors.

Estimating the optimal speed limit to maximize the traffic flow at a given density would be one important and popular topic in the future because traffic congestion has already increasingly influenced citizens' daily life. To obtain a more exact result at the optimal speed limits, we need to apply the simulation method to estimate the whole flow-density model for each sensor. And we can also apply the optimization method to solve the problem instead of just estimating the optimal speed limits. More meaningful works could be done in the future.

References

B.G. Heydecker, J.D. Addison, “Analysis and modelling of traffic flow under variable speed limits.” Transportation Research Part C, 19 (2) (2011), pp. 206-217

Boyd, S., Vandenberghe, L. (2004) Convex Optimization, University Press, Cambridge © Cambridge University Press, pp. 1, 67, 127, 338.

Caves, R. W. (2004). Encyclopedia of the City. Routledge. p. 141.

Cremer, M., 1979. “Der verkehrsfluss auf schnellstrassen: modelle, überwachung, regelung.” Springer-Verlag.

David Schrank and Bill Eisele (2019). “2019 Urban Mobility Report” Texas A&M Transportation Institute

Easa, S., M. (1982) “Selecting Two-Regime Traffic-Flow Models”, in Transportation Research Record Journal of the Transportation Research Board, pp. 35.

Edie, L. 1965. “Discussion of traffic stream measurements and definitions.” In Proceedings of the Second International Symposium on the Theory of Traffic Flow, edited by J. Almond, 139–154

FL Hall, K Agyemang-Duah 1991. “Freeway capacity drop and the definition of capacity.” Transportation research record No. 1320, Freeway Operations, Highway Capacity, and Traffic Flow 1991.

F. Soriguera, F. Robust. “Estimation of traffic stream space mean speed from time aggregations of double loop detector data.” Transp. Res. Part C Emerging Technol., vol. 19, no. 1, pp. 115-129, 2011

Greenshields, B. D. 1934. “A Study of Traffic Capacity.” Proceedings Highway Research Board 14: 448–477.

Hannah, L. A., & Dunson, D. B. (2013). Multivariate convex regression with adaptive partitioning. The Journal of Machine Learning Research, 14(1), 3261-3294.

Hegyi, A., De Schutter, B., Hellendoorn, H., 2005a. “Model predictive control for optimal coordination of ramp metering and variable speed limits.” Transp. Res. Part C 13 (3), 185–209.

Juho Ylimartimo. “Convex Quantile Regression for Traffic Congestion Modelling”. Mater Thesis, Business School of Aalto University, Fall 2019

Knoop, V. L., S. P. Hoogendoorn, and H. J. Van Zuylen. 2007. “Empirical Differences Between Time Mean Speed and Space Mean Speed.” In Proceedings of Traffic and Granular Flow 07, edited by C. Appert-Rolland, F. Chevoir, P. Gondret, S.Lassarre, J.-P. Lebacque, and M. Schreckenberg, 351–356. Paris: Springer.

Knoop, V. L., & Daamen, W. (2017). Automatic fitting procedure for the fundamental diagram. Transportmetrica B: Transport Dynamics, 5(2), 129-144.

Maerivoet, S., De Moor, B. (2005) Traffic Flow Theory, Physics. 1., pp. 10, 17-21.

Papageorgiou, M., Kosmatopoulos, E., Papamichail, I., 2008. “Effects of variable speed limits on motorway traffic flow.” Transport. Res. Rec. J. Transport. Res. Board 2047, 37–48.

R. C. Carlson, I. Papamichail, M. Papageorgiou, A. Messmer. “Optimal motorway traffic flow control involving variable speed limits and ramp metering.” Transp. Sci., vol. 44, no. 2, pp. 238-253, May 2010

R. Sen, A. Cross, A. Vashistha, V.N. Padmanabhan, E. Cutrell, W. Thies. “Accurate speed and density measurement for road traffic in India” Paper presented at the proceedings of the 3rd ACM symposium on computing for development (2013)

Soriguera F., Martínez I., Sala M., Menéndez M., “Effects of low speed limits on freeway traffic flow.” Tran. Res. C, 77 (2017), pp. 257-274

Varian, H. R. (1982). The nonparametric approach to demand analysis. Econometrica: Journal of the Econometric Society, 945-973.

Varian, H. R. (1984). The nonparametric approach to production analysis. Econometrica: Journal of the Econometric Society, 579-597.

Zackor, H. 1972. Beurteilung verkehrsabhängiger geschwindigkeitsbeschränkungen auf autobahnen. STRASSENBAU U STRASSENVERKEHRSTECH, (128).

Zackor, H. 1991. Speed limitation on freeways: Traffic-responsive strategies. In Concise Encyclopedia of Traffic & Transportation Systems (pp. 507-511). Pergamon.

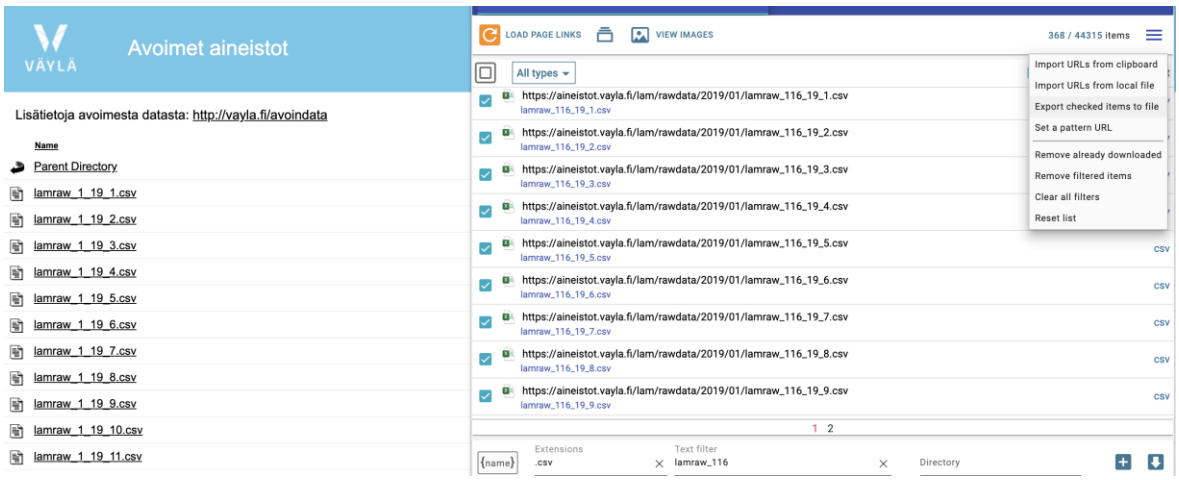
Appendix A: Appendix Title

```
# Figure 6.24, page 339.
# Least-squares fit of a convex function.

from cvxopt import solvers, matrix, spmatrix, mul
from pickle import load
solvers.options['show_progress'] = 0

data = load(open('cvxfit.bin','rb'))
u, y = data['u'], data['y']
m = len(u)

# minimize    (1/2) * || yhat - y ||_2^2
# subject to  yhat[j] >= yhat[i] + g[i]' * (u[j] - u[i]), j, i = 0,...,m-1
#
# Variables  yhat (m), g (m).

nvars = 2*m
P = spmatrix(1.0, range(m), range(m), (nvars, nvars))
q = matrix(0.0, (nvars,1))
q[:m] = -y

# m blocks (i = 0,...,m-1) of linear inequalities
#
#     yhat[i] + g[i]' * (u[j] - u[i]) <= yhat[j], j = 0,...,m-1.

G = spmatrix([],[],[], (m**2, nvars))
I = spmatrix(1.0, range(m), range(m))
for i in range(m):
    # coefficients of yhat[i]
    G[list(range(i*m, (i+1)*m)), i] = 1.0

    # coefficients of g[i]
    G[list(range(i*m, (i+1)*m)), m+i] = u - u[i]

    # coefficients of yhat[j]
    G[list(range(i*m, (i+1)*m)), list(range(m))] -= I

h = matrix(0.0, (m**2,1))

sol = solvers.qp(P, q, G, h)
yhat = sol['x'][:m]
g = sol['x'][m:]

nopts = 1000
ts = [ 2.2/nopts * t for t in range(1000) ]
f = [ max(yhat + mul(g, t-u)) for t in ts ]

try: import pylab
except ImportError: pass
else:
    pylab.figure(1, facecolor='w')
    pylab.plot(u, y, 'wo', markeredgecolor='b')
    pylab.plot(ts, f, '-g')
    pylab.axis([-0.1, 2.3, -1.1, 7.2])
    pylab.axis('off')
    pylab.title('Least-squares fit of convex function (fig. 6.24)')
    pylab.show()
```

https://cvxopt.org/examples/book/cvxfit.html

Appendix B: Simple Mass Downloader

A tool named Simple Mass Downloader is used to collect and download data. There is one example that showing how to download the yearly data of sensor 116 in the year of 2019. More details could be found via the link: https://gelprec.github.io/quick_start_v2.html



Appendix C: Code of downloading data from URLs file

```
import requests
import os
import pandas as pd
import glob

# download the csv file of each day from the txt URLs file
urlFIle = open("2019_116.txt", "r")

for line in urlFIle:
    url = line.strip('\n')
    filename = url[url.rfind("/")+1:]
    r = requests.get(url)

    with open(filename, 'wb') as f:
        f.write(r.content)


# combine multiple csv files of this year into one csv file
csv_list = glob.glob('*.csv')

# Calculate the number of csv files and make sure how many files missing
print(u'there are %s csv files'% len(csv_list))

for i in csv_list:
    fr = open(i,'rb').read()
    with open('2019_116.csv','ab') as f:
        f.write(fr)
```

Appendix D: The Abnormal Points